SLAM-based Dense Surface Reconstruction in Monocular Minimally Invasive Surgery and its Application to Augmented Reality

Long Chen[a], Wen Tang[a], Nigel W. John[b], Tao Ruan Wan[c], Jian Jun Zhang[a]

[a] *Bournemouth University*
[b] *University of Chester*
[c] *University of Bradford*

Abstract

Background and Objective While Minimally Invasive Surgery (MIS) offers considerable benefits to patients, it also imposes big challenges on a surgeon's performance due to well-known issues and restrictions associated with the field of view (FOV), hand-eye misalignment and disorientation, as well as the lack of stereoscopic depth perception in monocular endoscopy. Augmented Reality (AR) technology can help to overcome these limitations by augmenting the real scene with annotations, labels, tumour measurements or even a 3D reconstruction of anatomy structures at the target surgical locations. However, previous research attempts of using AR technology in monocular MIS surgical scenes have been mainly focused on the information overlay without addressing correct spatial calibrations, which could lead to incorrect localization of annotations and labels, and inaccurate depth cues and tumour measurements. In this paper, we present a novel intra-operative dense surface reconstruction framework that is capable of providing geometry information from only monocular MIS videos for geometry-aware AR applications such as site measurements and depth cues. We address a number of compelling issues in augmenting a scene for a monocular MIS environment, such as drifting and inaccurate planar mapping.

Methods A state-of-the-art Simultaneous Localization And Mapping (SLAM) algorithm used in robotics has been extended to deal with monocular MIS surgical scenes for reliable endoscopic camera tracking and salient point mapping. A robust global 3D surface reconstruction framework has been developed for building a dense surface using only unorganized sparse point clouds extracted from the SLAM. The 3D surface reconstruction framework employs the Moving Least Squares (MLS) smoothing algorithm and the Poisson surface reconstruction framework for real time processing of the point clouds data set. Finally, the 3D geometric information of the surgical scene allows better understanding and accurate placement AR augmentations based on a robust 3D calibration.

Results We demonstrate the clinical relevance of our proposed system through two examples: a) measurement of the surface; b) depth cues in monocular endoscopy. The performance and accuracy evaluations of the proposed framework consist of two steps. First, we have created a computer-generated endoscopy simulation video to quantify the accuracy of the camera tracking by comparing the results of the video camera tracking with the recorded

ground-truth camera trajectories. The accuracy of the surface reconstruction is assessed by evaluating the Root Mean Square Distance (RMSD) of surface vertices of the reconstructed mesh with that of the ground truth 3D models. An error of 1.24mm for the camera trajectories has been obtained and the RMSD for surface reconstruction is 2.54mm, which compare favourably with previous approaches. Second, *in vivo* laparoscopic videos are used to examine the quality of accurate AR based annotation and measurement, and the creation of depth cues. These results show the potential promise of our geometry-aware AR technology to be used in MIS surgical scenes.

Conclusions The results show that the new framework is robust and accurate in dealing with challenging situations such as the rapid endoscopy camera movements in monocular MIS scenes. Both camera tracking and surface reconstruction based on a sparse point cloud are effective and operated in real-time. This demonstrates the potential of our algorithm for accurate AR localization and depth augmentation with geometric cues and correct surface measurements in MIS with monocular endoscopes.

Keywords: SLAM, Surface Reconstruction, Augmented Reality, Minimally Invasive Surgery

1. Introduction

In Minimally Invasive Surgery (MIS), medical procedures are technically demanding, and the difficulty is exacerbated by well-known issues and restrictions associated with MIS, such as the limited field of view (FOV), lack of hand-eye alignment and orientation, and the lack of stereoscopic depth perception in monocular endoscopy. Augmented Reality (AR) technology can help overcome these limitations by overlaying additional information onto the real scene such as annotations at target surgical locations [18], labels [45], measurements of tumour sites [4] or even overlay a 3D reconstruction of anatomy [14] [15].

Despite recent advances in powerful miniaturized AR hardware devices and improvements on vision based software algorithms, many issues in medical AR remain unsolved. In particular, the dramatic changes in tissue surface illumination and tissue deformation as well as the rapid movements of the endoscope during insertion and extrusion, all give rise to a set of unique challenges that call for innovative approaches. As with any other technological assisted medical procedure, the accuracy of AR in MIS is paramount.

The miniaturized devices in MIS mean that the Field of View (FOV) captured by a monocular endoscopic camera is usually very small, for example, only 30% to 40% of the whole liver surface is visible in one frame at one time [38]. Traditional AR approaches (i.e. marker-less AR) for MIS are mainly based on feature tracking methods that require those selected feature points to be within the field of view [14]. Given the restricted FOV, the

Email addresses: chenl@bournemouth.ac.uk (Long Chen), wtang@bournemouth.ac.uk (Wen Tang), nigel.john@chester.ac.uk (Nigel W. John), t.wan@bradford.ac.uk (Tao Ruan Wan), jzhang@bournemouth.ac.uk (Jian Jun Zhang)

algorithmic limitations of traditional methods can severely affect the precision of AR for procedure guidance. Our proposed geometry-aware AR framework addresses the issue by providing global 3D geometric information of the entire surgical scene so that the information overlay does not depend on the frame by frame local feature extractions, hence, greatly improving the reliability of AR augmentations.

Studies have shown that a typical human uses 14 visual cues to perceive depth, and 11 of the 14 cues do not require binocular vision [11]. For example, depth information can be inferred in monocular vision through occlusions, motion parallax, shadows and texture gradient, and relative size and familiar size etc. The cognitive process of monocular vision enables surgeons to perform laparoscopic under a 2D environment [30]. However, monocular depth cues can only roughly estimate the general depth between objects, the accurate distance between objects cannot be perceived [41]. Although examples of stereoscopic endoscopes do exist, they are not commonly accessible in medical practice [47] [50]. We address the aforementioned challenges by providing accurate geometric measurements and artificially generating depth cues through AR technology, which are important improvements in monocular endoscope environment for surgeons to carry out complex procedures. In our AR framework, the distance between objects can be deciphered by relative sizes of AR labels and annotations.

A stereo endoscope can provide stereoscopic vision and such systems are currently available and often integrated into robotic systems (e.g. the da Vinci system from Intuitive Surgical, Inc.). 3D depth information can then be recovered using the disparity map from rectified stereo images during a laparoscopic surgery [42] [43] [8], so that a 3D reconstruction using a dense point cloud of the laparoscopic scene can be achieved by a propagation method [44] and/or a cost-volume algorithm [6]. Stereo vision based reconstructions, however, can only recover the structure of a local frame without a global overview of the scene, and are very sensitive to noise and luminance changes. Surgeons have to wear 3D glasses or use a binocular viewer on the robotic surgical system. In addition, stereo endoscopic surgery is still too expensive and yet to be widely used in practice. Hence, providing depth cues in monocular endoscope operations will have a significant impact on the accuracy of surgical procedures.

In this paper, we present a novel method and a computational framework to achieve accurate geometry-aware AR through: (i) extracting 3D depth information from camera motions and 3D surface reconstructions; and (ii) using AR technology to fuse rich 3D structural information with a monocular endoscope video stream, such that accurate spatial information in the scene can be derived from the scene geometry, and artificial depth cues can be provided based on the collaboration of the 3D spatial scene with the real-time video streams (i.e real-virtual overlay and simultaneous mapping). To this end, we explore the potential of the state-of-the-art SLAM framework by modifying and fine-tuning the algorithm for endoscopic camera tracking and mapping, so that the balance between point cloud density and computational performance can be achieved. A 3D surface reconstruction method based on the Moving Least Squares (MLS) smoothing and the Poisson surface reconstruction algorithms are proposed to recover a smooth surface from the unstructured sparse map points extracted from the MIS scene. Simulated laparoscopic sequences generated in a 3D

modelling package have been used to evaluate the performance of the proposed framework in terms of robustness of the camera tracking and the accuracy of the surface mesh reconstruction. Camera trajectories are compared with the ground truth camera trajectories, and the 3D surface reconstructions are measured against the 3D models of the simulated laparoscopic scene. The experimental results yield root mean square errors (RMSE) of 1.24 mm for camera trajectories and 2.54mm for the surface reconstruction.

The obtained global geometric information can be seamlessly integrated into our proposed AR framework, which is capable of achieving AR augmentations at the correct depth and detailed accurate surface measurements. Our method provides new possibilities for novel geometrically informed AR augmentations in monocular endoscopic MIS, including accurate annotations, labels, tumour measurement and artificial depth cues at correct depth locations that are demonstrated with two example applications: i.e. generations of artificial depth cues and the surface measurements of target sites in MIS.

2. Previous work

Recent advances in computer hardware and software technologies have facilitated the use of computer vision techniques for MIS scene guidance and information augmentation. For example, AR guidance systems have been used to visualize pre-operative CT images [18] [45], for tumour AR visualization in laparoscopic surgery [4] and anatomy structures AR mapping in liver MIS surgery [14] [15]. There are, however, some particular challenges faced with AR in MIS. The luminance changes dramatically and an endoscope can move rapidly during insertion and extrusion. Traditional tracking methods for AR in MIS usually involve feature points based tracking such as Scale-Invariant Feature Transform (SIFT) [18], Speeded Up Robust Features (SURF) [22], Optical Flow tracking [38] or other approaches specifically designed to work with soft tissues that account for changes in scale, rotation and brightness [31]. As these invariant descriptors are designed for 2D tracking, the information regarding the depth within a scene has not been recovered and selected feature points extracted from vision algorithms must be within the field of view, resulting in the lack of global information in AR augmentations.

Constraint-based factorization methods (CBFM) [51] provided a computational solution for 3D structure reconstructions from 2D endoscopic images, but external tracking devices are needed to provide the surgical instruments position, reducing its usability in practice. Additional work has also investigated using shadows for inferring the depth from monocular endoscopic images [48], but these shape from shading methods rely on the strong assumption that a single point illumination source is being used without any reflection and can also be affected by different tissue colours. Whereas in real laparoscopic surgery, the diffuse and specular reflection does exist due to the complex surface conditions of different tissues. This will severely affect the accuracy of shape form shading, and we compared our reconstruction results with shape form shading in Section 4. Lin *et al* [25] combined structured lighting with structure from motion for monocular endoscopic image reconstruction. Although special optical probe is needed, better reconstruction density and robustness are achieved with extra benefit of super spectral resolution. With the recent development of deep learning

technology, it is possible to use Convolutional Neural Networks (CNN) for estimating depth from a single endoscopic image [49]. However, the inference results of the network highly depends on the dataset used for training; it is still very difficult to build a large dataset with groundtruth for surgical scenes.

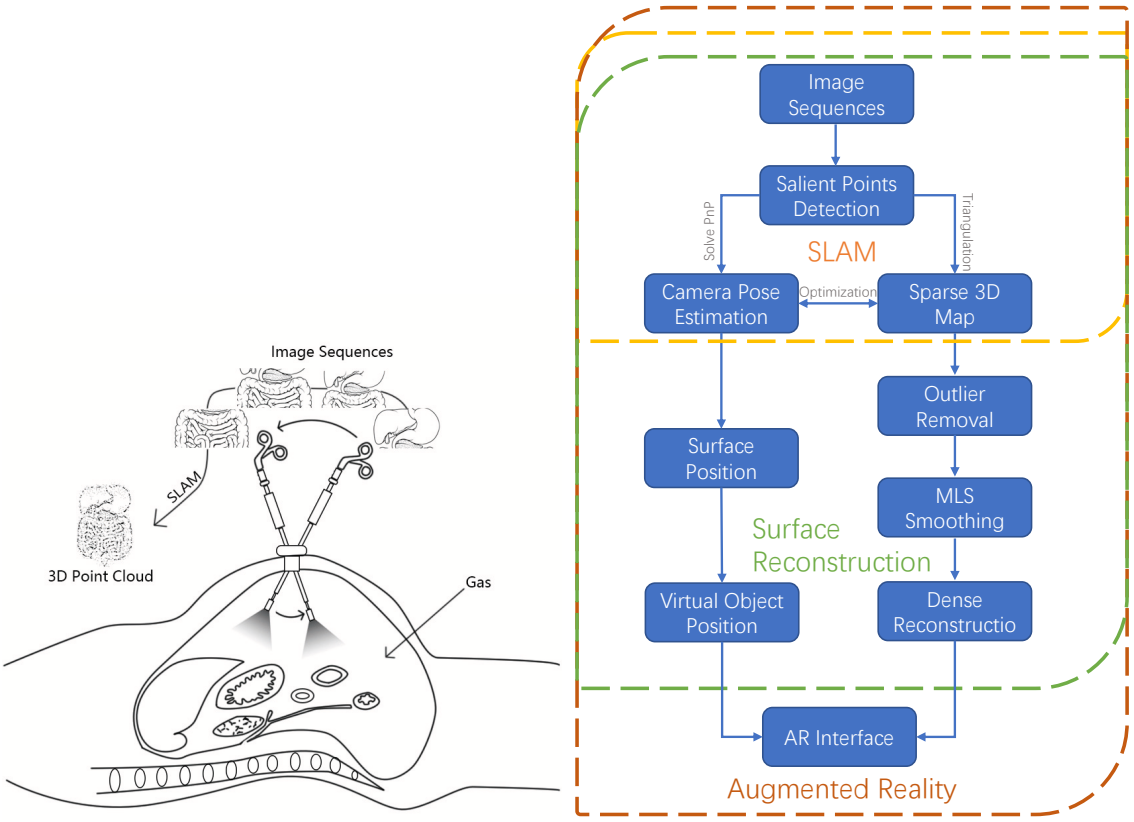
Recently, the maturity of the method of simultaneous localization and mapping (SLAM) designed for robot navigation in unknown 3D environments has opened up new opportunities for developing novel endoscopic camera tracking approaches in MIS. SLAM-enabled systems make it possible to estimate the 3D structure of the MIS scene from a moving endoscope camera and simultaneously track the pose of the camera. The scenario of the camera tracking and scene reconstruction in endoscopic surgeries is similar to that of a typical SLAM application in robotic vision, albeit with additional challenges. SLAM based approaches do not require the use of optical or magnetic trackers and have been tested for camera tracking with laparoscopic image sequences [32] [13] [5] [2]. A motion compensation model [33] and the stereo semi-dense reconstruction method [46] have also been integrated into the EKF-SLAM framework to deal with the tissue motion problem. However, due to the linearization of the motion and sensor models by a first-order Taylor series expansion, the accuracy of the EKF-SLAM cannot be guaranteed, thus, it is prone to inconsistent estimation and drifting in the camera motion estimation. The PTAM (Parallel Tracking and Mapping) algorithm was a breakthrough in visual SLAM and has been also used in MIS for stereoscope tracking [24]. Derived from the PTAM, ORB–SLAM (Oriented FAST and Rotated BRIEF) is a well-designed SLAM system that utilities ORB binary features for a fast and reliable feature point tracking. Mahmoud *et al* [28] tested ORB–SLAM on endoscope videos and presented a method for densifying map points, but the algorithm has some loss of accuracy.

3. Methodology

The flowchart in Figure 1 demonstrates our intra-operative MIS AR framework. As can be seen from Figure 1 (a), the endoscope is inserted into the patient abdominal cavity, which is inflated with carbon dioxide gas to create the pneumoperitoneum. Image sequences captured by the moving endoscopic camera are the input to our AR framework as shown in Figure 1 (b). The SLAM algorithm recovers the camera pose and generates an unorganized sparse point cloud. 3D geometric information is then built based on the point cloud by our proposed surface reconstruction framework. The dense surface mesh is then aligned with the input image sequences via a camera space transformation. Finally, the virtual object can be displayed onto the reconstructed surface to provide both depth cues and any virtual augmentation at the correct depth.

3.1. Introducing of the surface coordinate

The difference between our approach and the feature based tracking method used in the previous AR work in MIS [18] [15] is shown in Figure 2, illustrating the use of different coordinate systems. The endoscope is represented as a probe in the camera coordinate system and p_c is a 2D point in the camera's view and the virtual object (the face) to be displayed. Figure 2 (a) shows the feature tracking based AR environment. When the feature



(a) (b)

Figure 1: (a). A moving monocular endoscopic camera can capture a series of image sequences which can be used to build a 3D sparse point cloud by using a SLAM system. (b) The flowchart of our proposed AR framework.

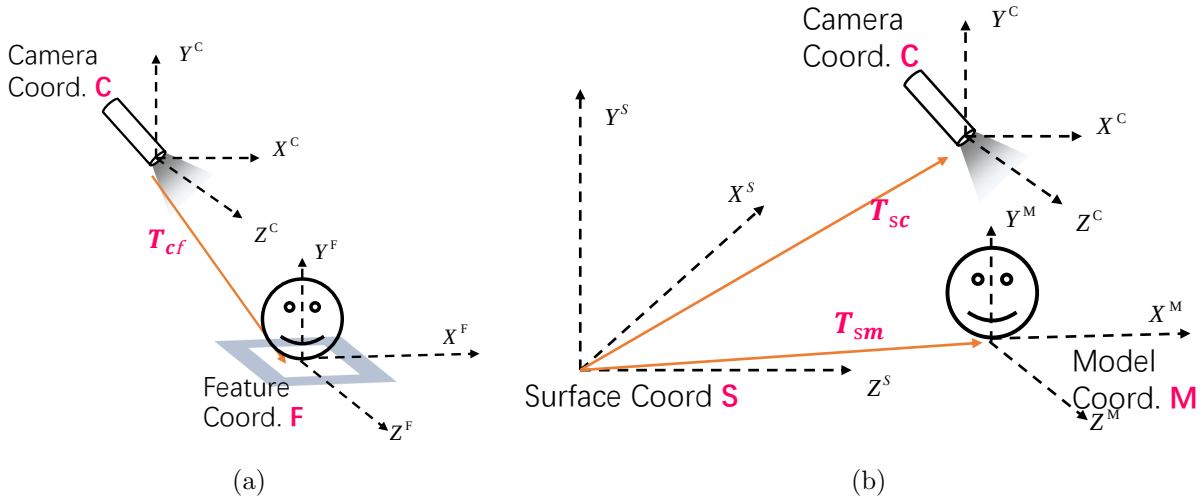


Figure 2: The comparison of the marker-less tracking (a) and our proposed AR framework (b).

is detected and tracked, the virtual object will be placed in the feature coordinate system. Assuming P_f is a 3D point in the MIS scene, then the 3D point can be transformed to the 2D point in the endoscopic camera's view by the following equation:

$$p_c = K * T_{cf} * P_f \tag{1}$$

where T_{cf} can be computed by solving the Perspective-n-Point (PnP) problem, and K is the camera intrinsic parameters. For our proposed AR framework, as can be seen from Figure 2 (b), we add a surface local space S as an agent, which serves as the intermediary and is incrementally built from the point cloud sensed in the environment, which allows us to achieve great robustness. A 3D point in the model space P_m can be transformed to the 2D camera space p_c by:

$$p_c = K * T_{sc}{}' * T_{sm} * P_m \tag{2}$$

where T_{sc} is estimated by the SLAM and T_{sm} is a user-defined matrix. By using the local surface space as an agent, we solved two important issues for AR in MIS: (i) no pre-captured or manually selected features are needed, which saves time and enables 360 degree tracking; (ii) AR objects can be placed anywhere on the surface at the correct depth.

3.2. Monocular endoscopic camera tracking and mapping

We use ORB–SLAM [35], which outperforms many SLAM systems such as Mono-SLAM [10], PTAM [19] and LSD-SLAM [12], for the task of monocular endoscopic camera tracking and mapping. ORB–SLAM combines many state-of-the-art techniques into one SLAM system, such as using an ORB descriptor [40] for tracking, local keyframe for mapping, graph-based optimization, the Bag of Words algorithm for re-localization, and an essential graph for loop closure. These features can enable real-time endoscopic camera tracking and sparse point mapping in an abdominal cavity as shown in Figure 1. Real-time performance

is crucial in time-demanding medical interventions. Since ORB [40] is a binary feature point descriptor, it is an order of magnitude faster than SURF [1] and more than two orders faster than SIFT [27] with better accuracy. In addition, ORB features are invariant to rotation, illumination and scale, which means that it is capable of dealing with some of the main challenges in MIS scenes including rapid movements of endoscope cameras (rotation and zooming) and the change of brightness.

Initialization

A common problem for monocular scene analysis using SLAM is the initialization, a step required for generating an initial map, because the depth cannot be recovered from a single image frame. An automatic approach is used in ORB–SLAM to calculate homography for planar scenes and a fundamental matrix for non-planar scenes dynamically. This approach can greatly increase the success rate of initialization and reduce the time required for the initialisation step. It also facilitates the initialization on an organ surface or to compute a fundamental matrix when the endoscopic camera is pointing at complex structures.

Training of data sets

One of the huge challenges that is unique to AR in MIS is the rapid movement of endoscopes due to constant extraction and insertion of the device. The tracking algorithm must be robust to accommodate the loss of image sequences after an extraction, and recover the tracking during a re-insertion. The Bags of Words (BoW) algorithm solves this re-localization problem during the tracking. In the BoW algorithm, the vocabulary is created offline with a large number of ORB descriptors extracted from very large data sets of images that cover almost all of the patch patterns that may be encountered. The vocabulary serves as a classifier or a dictionary to assign each descriptor an index. When a new image appears in the system, each descriptor of features in this image is looked up, and a unique vector will be built based on the index of descriptors. In doing so, the rough similarity of two images can be acquired by simply comparing the two unique vectors, therefore, it can greatly increase the speed of re-localization.

The default BoW database in ORB–SLAM contains a very large image data set with different genres captured from the real world scenes. Such a universal database would be too sparse and general for specific MIS tasks. When processing endoscopic videos, images are generally captured inside of human bodies for different organs, tissues and vessels. These MIS scenes are more homogeneous and specific than the real word scenes. Therefore, we trained our vocabulary list specifically for its use in MIS based on 877 images sampled from ten *in vivo* sequences obtained from the Hamlyn Centre Endoscopic Video database [26] [52]. By training a specific MIS BoW database, the specific features existing in the minimally invasive surgery scenes are collected and saved. The length of the unique vectors for similarity measurements will be decreased, hence, reducing not only the loading time of the AR framework, but also the time of BoW query as shown in Table 1.

[1]Based on the average time of 1000 times' BoW query experiment

Table 1: Comparison of the original and our trained BoW database

Item	***Original BoW Database***	***Database Trained for MIS***
Training Source:	Images with Different Genres	Endoscopy Videos [52]
Database Size:	145.3 MB	41.8 MB
Number of Words:	971815	259677
BoW Query Time[1]:	4.85ms	4.42ms

This approach generalizes well to different MIS scenes since the training based on the Hamlyn Centre Endoscopic Video Database covers a range of medical scenarios from gastrointestinal examinations, diaphragm dissection, lung lobectomy, coronary artery bypass, to cardiac examination.

Parameter tuning and increasing surface points

We fine-tuned some of the parameters that were used by default in the ORB–SLAM by increasing the limit of the number of features extracted per image by a factor of two, which allows a maximum of 2000 feature points to be detected. The maximum threshold that is allowed between keypoints and reprojected map points for triangulation is reduced by a factor of ten to constrain the range of the points to be selected so that strictly robust 3D points are chosen and feature points moved by tissue deformation are rejected. This approach can greatly improve the tracking accuracy. Finally, the Hamming distance threshold for the ORB descriptor comparisons is decreased by 0.8 for more strict applications of the pair point rule. After tuning the default parameters, around 50% more map points can be detected for the reliable surface reconstruction pipeline. Furthermore, the system has the ability to filter small drifts caused by tissue deformations with strict map point selection criteria.

3.3. Intra-operative 3D surface reconstruction

One of the main advantages of our proposed AR system is its ability to use a sparse 3D point cloud extracted from a moving monocular endoscopic camera to construct a dense and smooth surface through our novel surface reconstruction framework. Our framework processes the unstructured sparse point clouds using a combination of outlier removal filters, the Moving Least Square algorithm to smooth noise data and a Poisson surface reconstruction method to generate the smooth surface from an unstructured sparse point cloud. This pipeline is illustrated in Figure 3. Details of each processing step are presented in the following sections.

Point cloud pre-processing

The point cloud P given by ORB–SLAM represents salient points that are visible at different camera keyframes, giving a sparse representation of the intra-operative scene. MIS scenes are very complex due to issues associated with camera calibrations and movements and reflections of tissues. Hence, the result is a noisy point cloud mixed with many outliers that can affect the final surface reconstruction. Our approach to solve this problem is to

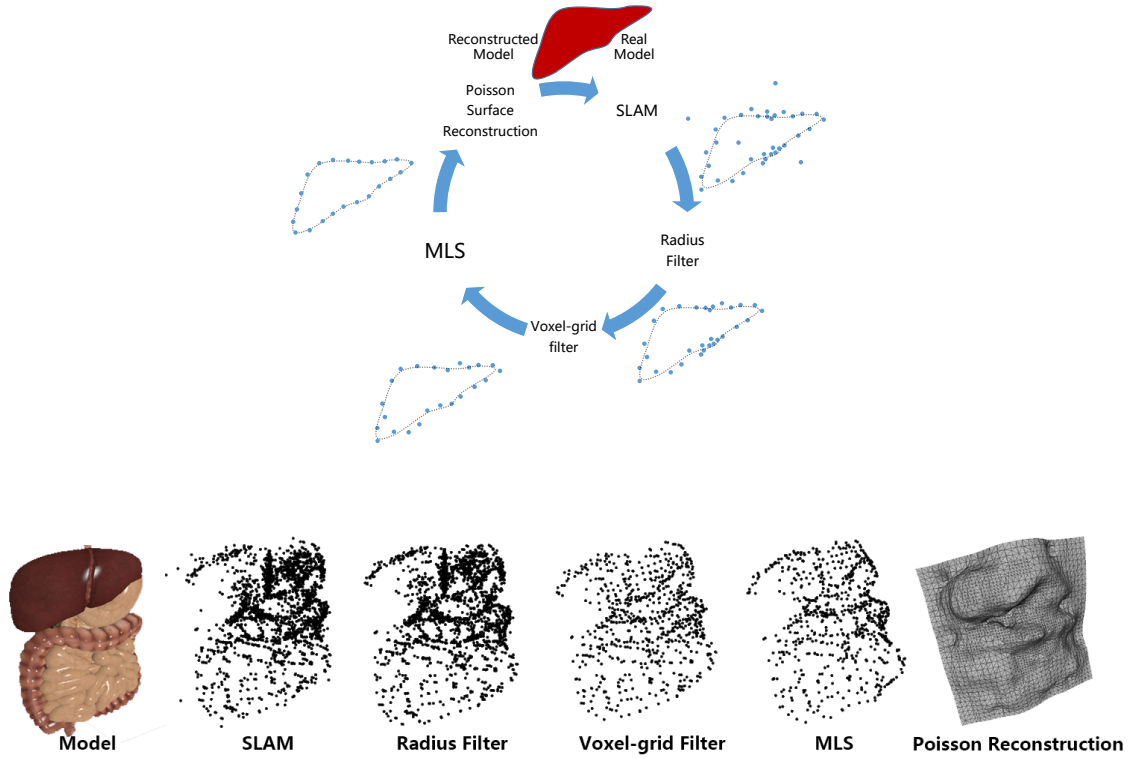


Figure 3: The proposed intra-operative 3D surface reconstruction framework.

apply two outlier removal filters to remove the noisy points located amongst the raw data points before feeding the point cloud into the reconstruction pipeline.

Firstly, a radius filter is used to process points in a cloud based on the number of neighbour points. Points with very few neighbours are labelled as outliers or isolated points that should not contribute to the overall structure of the 3D scene. Since some texture-abundant areas gain many more points than other areas, a voxel-grid filter is then used to re-sample the point cloud into a more evenly distributed point cloud. After the filtering process, the point cloud becomes 'clean' and ready for MLS (Moving Least Square) smoothing and 3D surface reconstruction.

Moving Least Square point smoothing

The Moving Least Squares (MLS) algorithm [23] reconstructs surfaces locally by solving an optimization problem to find a local reference plane and fit a polynomial to the surface. Let a point set $p_i \in \mathbb{R}_3, i \in \{1, ..., N\}$ be the point cloud produced from the ORB–SLAM system. the continuous and smooth MLS surface S can be computed by a two-step procedure: (i) a local reference plane is defined as $H = \{x \in \mathbb{R}_3 | x \cdot n - D = 0\}$, which can be computed by minimizing the weighted sum of squared distances:

$$\sum_{i=1}^{n} (p_i \cdot n - D)^2 \Phi(\|p_i - q\|)$$

where q is the projection of p onto H, and Φ is the MLS kernel, usually a Gaussian; (ii) after the points are projected onto the initial local reference plane, a second least squares optimization is used to find a bi–variate polynomial function $g(u, v)$ (where u, v is the local coordinate of q in H) that approximates to the local surface. The projection of p onto S can then be defined by the polynomial value at the origin, i.e. $q + g(0, 0) \cdot n$.

Poisson surface reconstruction

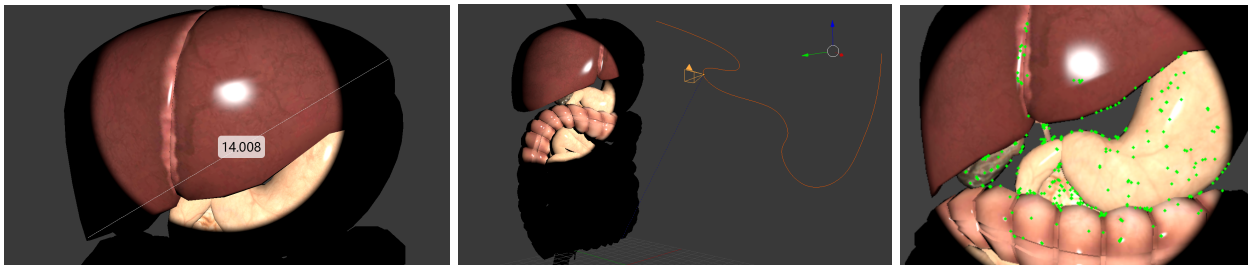
We represent the points after the MLS filter stage by a vector field $\overrightarrow{V}$. Poisson surface reconstruction [29] approaches the surface reconstruction problem through a framework of implicit functions that compute a 3D indicator function χ (which is equal to 1 inside the model and 0 at the outside points). Therefore, the problem becomes that of finding the χ whose gradient is the best approximation of the vector filed $\overrightarrow{V}$:

$$min_\chi \left\| \bigtriangledown_\chi - \overrightarrow{V} \right\|$$

Applying the divergence operator, we can transform this into a Poisson problem:

$$\bigtriangledown \times (\bigtriangledown_\chi) = \bigtriangledown \times \overrightarrow{V} \Leftrightarrow \triangle_\chi = \bigtriangledown \times \overrightarrow{V}$$

After solving the Poisson problem and obtaining the 3D indicator function χ, the 3D surface can be directly obtained by extracting an isosurface [17]. The Poisson reconstruction process acts as a global solution that treats all of the data points simultaneously without relying on a heuristic partitioning or blending, so that it can robustly approximate noisy data and create very smooth surfaces.



(a) (b) (c)

Figure 4: Simulated MIS scenes with a realistic human digestive system model. (a) The size of the model is scaled to the real world size of an adult liver. (b) The only light is attached to the camera and the camera trajectory is designed to hover around the 3D model. (c) The frame that ORB–SLAM succeeded in initializing.

4. Results

We designed a two-part quantitative and qualitative evaluation process: (i) using a realistic simulation of a MIS scene video for the ground truth study to assess the performance of the SLAM tracking error and the accuracy of the proposed surface reconstruction framework; (ii) using a real *in vivo* video acquired from the Hamlyn Centre Laparoscopic/Endoscopic Video Datasets [26] [34] to assess the quality of our proposed framework.

4.1. System setup

Our system is implemented in an Ubuntu 14.04 environment using C/C++ (without any GPU acceleration). All experiments are conducted on a workstation equipped with Intel Xeon(R) 2.8 GHz quad core CPU, 32G Memory, and one NVIDIA GeForce GTX 970 graphics card. The size of the simulation image sequences is 1024 X 768 pixels and the size of *in vivo* endoscope video is 840 X 640 pixels. ORB–SLAM with our proposed AR framework runs in real-time at 40 FPS at max and the 3D surface reconstruction process takes around 600ms to traverse the whole pipeline.

4.2. Ground truth study using simulation data

For the evaluation of the accuracy of tracking performance, all camera trajectories estimated by ORB–SLAM were aligned with trajectories of the ground truth camera used to render the MIS scene video. Similarly, the accuracy of our proposed 3D surface reconstruction framework is evaluated by comparing the reconstructed surface with the 3D model used to render the simulation video.

To quantitatively evaluate the performance of ORB–SLAM, we used Blender [3] – an open source 3D creation software to render realistic image sequences of a simulated abdominal cavity scene using pre-defined endoscopic camera movements. The digestive system contains 3D models with textures to make the scene as realistic as possible. The model was scaled to be the real life size according to an average measured liver diameter of 14.0 cm [21] as shown in Figure 4(a), the material property was set with a strong specular component to

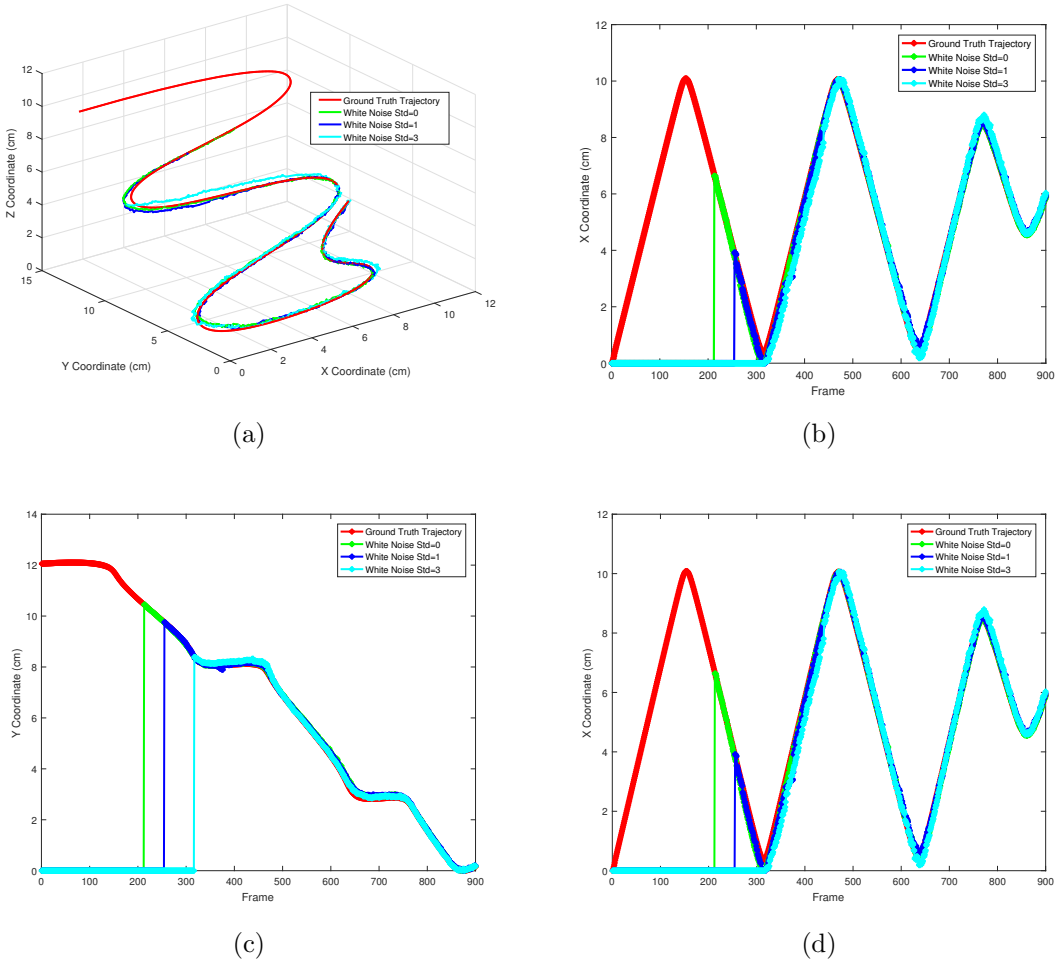


Figure 5: The camera trajectory comparison of the ground truth (red dots) with the estimated results under different white noise levels: no white noise (green dots), white noise SD=1 (dark blue dots), and white noise SD=3(light blue dots) in four different views, (a) 3D view, (b) view of X-axis, (c) view of Y-axis, (d) view of Z-axis

simulate the smooth and reflective liver surface tissue. The luminance is intentionally set high with a spot light attached to the main camera to simulate an endoscope camera as shown in Figure 4(a) to render a realistic endoscopic lighting condition. We designed a camera trajectory that hovers around the 3D model (Figure 4(b)) to capture as much of the area as possible so as to build a point cloud that could cover the whole front surface of the models. Nine hundred frames of image sequences were captured at a frame-rate of 30 fps, which is equivalent to a 30 second video. In order to investigate the robustness of our framework, we intentionally add white noise with different standard deviation (SD) to the synthetic video. We now have three version of the synthetic videos (with no white noise, white noise SD=1, and white noise SD=3, respectively), which will together be used for the further evaluation.

Camera trajectory evaluation

Fig 4(c) shows one of the rendered images from the sequences used as the input to ORB–SLAM. The camera trajectory started with a close shot location of the liver surface. ORB–SLAM was successfully initialized around frame 200 to 300 when the camera was in a place and where many feature points were identified. After the initialization step, the SLAM system ran stably and the camera trajectory was estimated with the origin of the coordinate system at the initialized position. The estimated camera trajectory was then extracted and normalized into the same coordinate system as that of the simulated ground truth model to assess the SLAM tracking performance.

Figure 5 shows the performance evaluation results; Figure 5(a) displays the camera trajectories in 3D space, in which green, dark blue and light blue dots represent the camera trajectory estimated by ORB–SLAM under no white noise, white noise SD=1 and white noise SD=3. Red dots are the trajectory of the simulated ground truth. Figs. 5(b), (c), and (d) shows the camera trajectories in X-axis, Y-axis, and Z-axis views, respectively. As can be seen, the SLAM camera trajectory starts at frame 212, 254 and 316 for the video with no white noise, white noise SD=1 and white noise SD=3, respectively, as there is no estimated data before initialization. Once the camera tracking is initialized, the trajectory of the camera matches closely with the ground truth camera trajectory represented by red dots. RMSE between the two camera trajectory data sets was also calculated with results of 1.24mm, 2.33mm and 4.39mm.

3D surface reconstruction evaluation

When the ORB–SLAM system gained enough feature points, we build a 3D surface based on the sparse point cloud. The whole reconstruction pipeline takes only 600 ms to generate the surface, which was then exported into the 3D model space to be compared with the ground truth surface data set. A simple iterative closest point (ICP) algorithm was used to align the reconstructed surface with the 3D model that was used to render the video. Root Mean Square Distance (RMSD) is used to evaluate the overall distance between the two surfaces. They are aligned in the real world coordinate system and we apply a grid sample to get a series of x,y coordinate points based on the surface area, and then compare the distance of the z value of the two surfaces.

$$RMSD = \sqrt{\frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n}(Z_{x,y} - z_{x,y})^2}$$

The RMSD to the ground truth surface is 2.54mm, 2.81mm and 3.66mm for the surface reconstructed by our proposed framework with different white noise levels. As shown in Table. 2, our proposed method is much more accurate than the Shape-from-Shading (SFS) [39][48] method as SFS was based on the strong assumption of a single point illumination source and can be affected by different tissue colours. Also, the reconstruction error of our method is better than that of the sparse cloud points reported as 4.10mm [28][35]. Also, our method can reconstruct a dense surface compared to that of the less clinical applicable sparse method. To further evaluate our reconstruction result, we have also rendered our

Table 2: Surface reconstruction results

Type	**Method**	**RMSD(wn=0)**	**RMSD(wn=1)**	**RMSD(wn=3)**
Mono/Dense	SFS[39][48]	7.21mm	8.38mm	11.60mm
Mono/Dense	**Proposed**	**2.54mm**	**2.81mm**	**3.66mm**
Stereo/Dense	BM [9]	2.04mm	2.09mm	2.17mm
Stereo/Dense	Chang *et al*[6]	2.57mm	2.21mm	2.28mm

video in stereo mode and tested it with popular stereo reconstruction approaches such as Block Matching (BM) and the state-of-the-art cost volume stereo reconstruction method by Chang *et al*. Our method is slightly better than the cost volume when there is no white noise, but overall is less accurate than stereo reconstruction, as the depth can be directly calculated from the disparities of stereo image pairs.

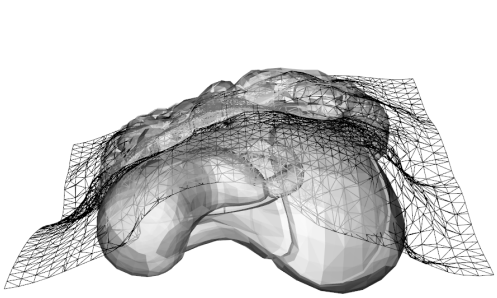
Figure 6 (a) shows that the reconstructed 3D surface aligns with the 3D model closely; Figure 6 (b) shows the top down view of the alignment. Figure 6 (c) shows the distance map between the reconstructed surface with the 3D ground truth model, where warm colours show penetrations between the two surfaces, the green colour represents a perfect match between the two surfaces, and the blue colour shows the largest distance between the two surfaces.

4.3. Real endoscopic video evaluation

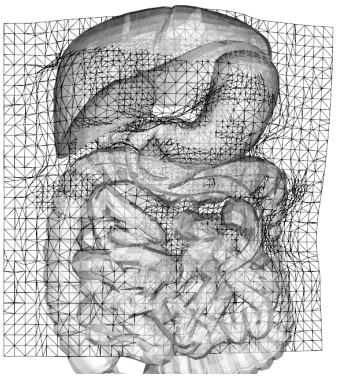
To qualitatively evaluate the performance of our proposed surface reconstruction framework, we applied the proposed approach with the real *in vivo* videos from the Hamlyn Centre Laparoscopic / Endoscopic Video Datasets. Figure 7 (a) (e) and (f) shows the reconstruction results from our 3D reconstruction framework. Figure 7 (b) shows the depth augmentation by fusing the camera pose from the SLAM system and the 3D surface reconstructed from our proposed framework. The real-time alignment of the 3D transparent mesh and the video are a good match, suggesting that our method can provide the correct depth information intra-operatively and so help improve surgical performance by displaying 3D mesh structures when performing monocular endoscope procedures. However, when large deformation occurs or the surgical instruments occupy the large proportion of the view, our framework may fail as shown in Figure 7 (f).

With our new 3D surface reconstruction approach, we have developed a geometry-aware AR framework for depth correct AR argumentation within the intra-operative endoscope scene in real-time. Our AR framework is an important step towards high quality AR in MIS, since incorrect depth placement will cause virtual objects to appear to drift away when the viewing angle changes. Furthermore, accurate global geometric information plays a crucial role in augmenting the real surgical scenes with annotations, labels, tumour measurements, inguinal measurements to estimate optimal mesh size for inguinal herniorrhaph [20] or even a 3D reconstruction of anatomy structures at the target surgical location. We demonstrate two example applications to show the clinical relevance.

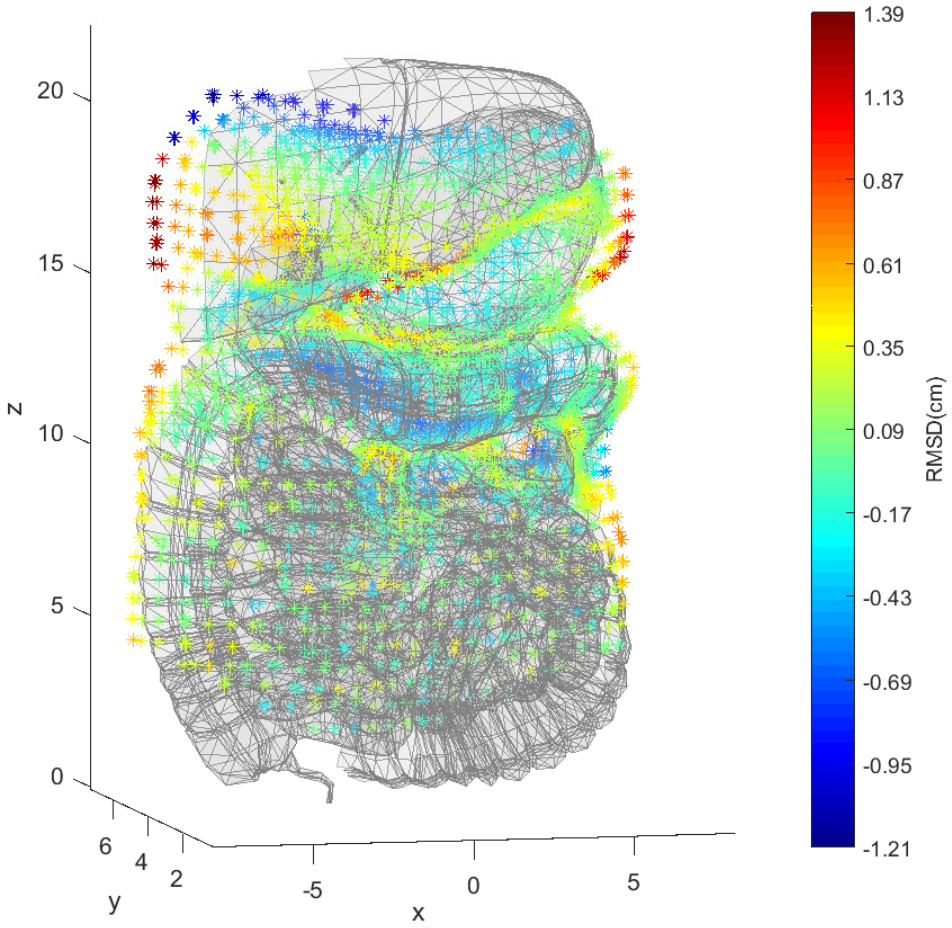
In Figure7 (a), AR augmentations of 3D arrows labels are placed onto the video frames to generate artificial depth cues and Figure7 (b) shows that virtual 3D arrows exist at different depths within this geometry-aware environment. In the second example, we recover the scale



(a)

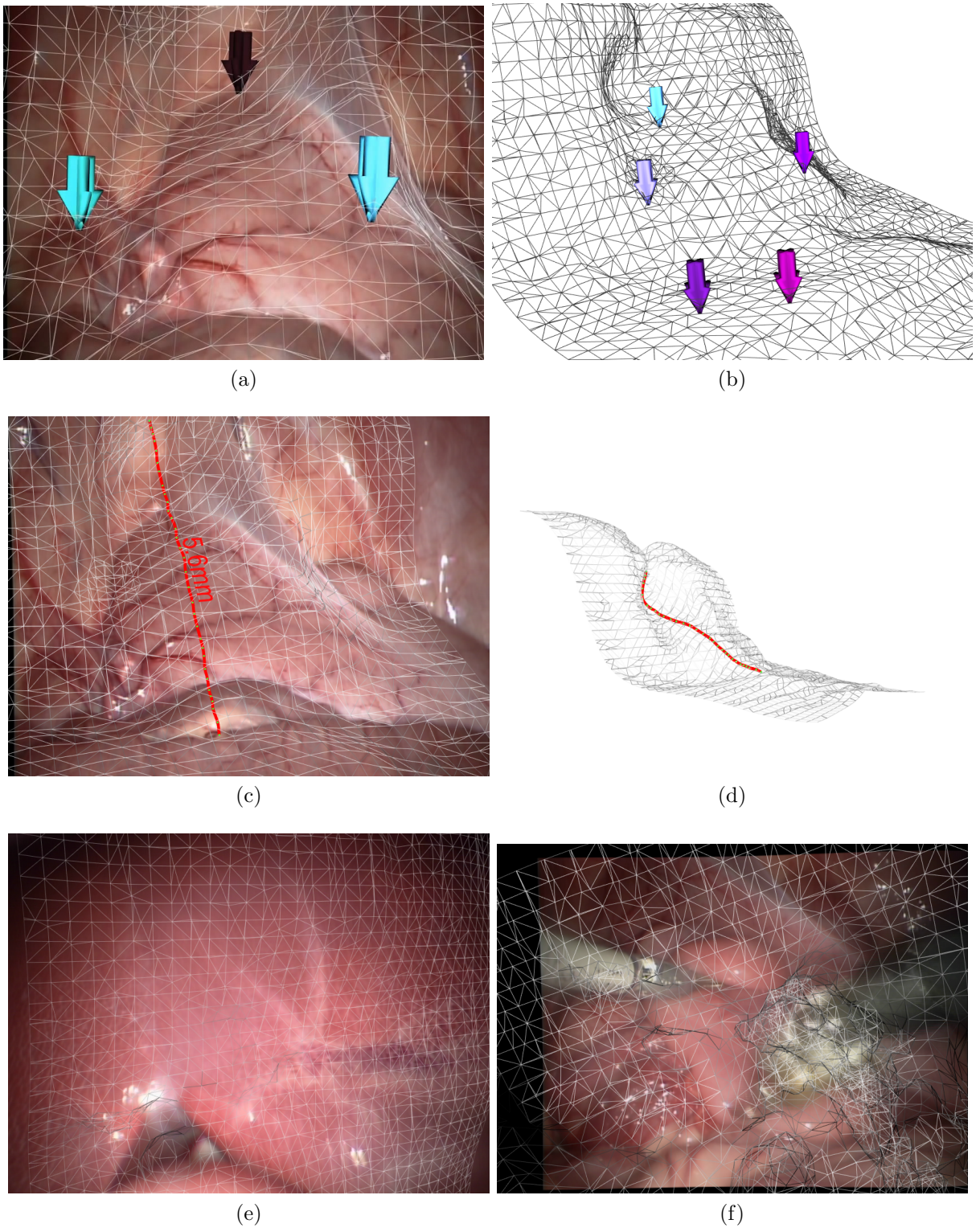


(b)



(c)

Figure 6: (a) and (b): the surface nicely represents the model surface. (c) Surface reconstruction error map.



(a) (b) (c) (d) (e) (f)

Figure 7: The surface reconstruction results applied to an *in vivo* video sequence. (a) Interactively adding arrows as annotations intra-operatively.(b) The view of mesh to show the annotations are in different depth.(c) Intra-operative measurement example. (d)The side-view of the intra-operative measurement example. Note that the measurement line follows the surface curvature closely. (e) The augmented mesh on a liver. (f) Our framework may fail when large deformation occurs or the surgical instruments occupy large proportion of the view

to the real-world size [37] to enable accurate intra-operative measurement as demonstrated in Figure7 (c) and (d). Note that measurement (the red line) follows the surface curvature closely, providing accurate results with correct depth information. More details can be appreciated in our demonstration video [7].

5. Discussion

Intra-operative MIS scene reconstruction is a challenging task especially for monocular MIS scene that the only input source is the monocular video stream. Acquiring the depth and geometric information in MIS is crucial for not only AR tasks such as intra-operative measurement, but also enables the potential applications of skill evaluation [16], autonomous tasks such as autonomous ultrasound scanning [53], debridement and cutting [36]. We are able to achieve a promising reconstruction result by our proposed SLAM-based monocular reconstruction approach (RMSD = 2.54mm), which is much accurate than other monocular MIS scene reconstruction method (RMSD = 7.21mm) and even comparable to the state-of-the-art stereo reconstruction method (RMSD = 2.04/2.57mm) that the depth can be directly derived from the disparity of stereo vision.

The limitation of our proposed method is that the SLAM theory is developed based on static world assumption; the deformations of objects (such as tissues and organs) directly challenge this basic condition for SLAM to estimate camera poses for 3D reconstruction. Therefore, soft tissue deformation is a great challenge to support in the SLAM based reconstruction framework as proposed here. Particularly with monocular endoscopic videos, it is extremely hard to recovery the soft deformation correctly while simultaneously estimating the camera poses. For small deformations like those in the *in-vivo* video that we use, however, the RANSAC algorithm in SLAM system will filter the outliers and recover the correct movement. For large deformation in very small FOVs, it is still unclear how to solve the tissue deformation issue without using extra external sensors within the monocular scene.

Using stereoscopic views is a possibility and we will investigate this in future work. One possible solution to accurately simulate and track the deformation is to to use real-time deformation model[54] and feature-based tracking [22] to recovery the movement of tissue. Although the accuracy and speed of our framework are acceptable, we will continue developing a dense SLAM system to be used in MIS reconstruction and extend the current reconstruction framework. This will enable us to develop a prototype system that can be tested in the operating theatre with our clinical collaborators, further investigating the benefit and efficacy of our approach and providing evidence for our hypothesis that visual SLAM can enhance the tools available to surgeons performing monocular endoscopic procedures.

6. Conclusions

In this paper, we presented an efficient and effective 3D surface reconstruction framework for an intra-operative monocular laparoscopic scene based on SLAM. This new approach has shown promising results when tested on both simulated laparoscopic scene image sequences and clinical data. The proposed framework also reveals several potential clinical applications such as additional depth cues augmentation and geometry-aware augmented reality in MIS.

[1] Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: Computer vision–ECCV 2006, Springer, pp 404–417
[2] Bergen T, Wittenberg T (2016) Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods. IEEE journal of biomedical and health informatics 20(1):304–321
[3] Blender (2016) Blender - free and open 3d creation software. URL https://www.blender.org/, [Accessed 6 Nov. 2016]
[4] Bourdel N, Collins T, Pizarro D, Debize C, sophie Grémeau A, Bartoli A, Canis M (2017) Use of augmented reality in laparoscopic gynecology to visualize myomas. Fertility and Sterility DOI 10.1016/j.fertnstert.2016.12.016
[5] Burschka D, Li M, Ishii M, Taylor RH, Hager GD (2005) Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery. Medical Image Analysis 9(5):413–426
[6] Chang PL, Stoyanov D, Davison AJ, Edwards PE (2013) Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. Med Image Comput Comput Assist Interv 16(Pt 1):42–49
[7] Chen L (2016) Supplemental video. URL https://youtu.be/Y9D3Liw5tXo, [Accessed 12 Feb. 2017]
[8] Chen L, Tang W, John NW (2017) Real-time geometry-aware augmented reality in minimally invasive surgery. Healthcare Technology Letters URL http://digital-library.theiet.org/content/journals/10.1049/htl.2017.0068
[9] Chen YS, Hung YP, Fuh CS (2001) Fast block matching algorithm based on the winner-update strategy. IEEE Transactions on Image Processing 10(8):1212–1222, DOI 10.1109/83.935037
[10] Davison AJ, Reid ID, Molton ND, Stasse O (2007) Monoslam: Real-time single camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6):1052–1067, DOI 10.1109/TPAMI.2007.1049
[11] Dunkin BJ, Flowers C (2015) 3d in the minimally invasive surgery (mis) operating room: Cameras and displays in the evolution of mis. In: Imaging and Visualization in The Modern Operating Room, Springer, pp 145–155
[12] Engel J, Schps T, Cremers D (2014) LSD-SLAM: Large-scale direct monocular SLAM. In: Computer Vision – ECCV 2014, Springer Nature, pp 834–849
[13] Grasa OG, Bernal E, Casado S, Gil I, Montiel J (2014) Visual slam for handheld monocular endoscope. Medical Imaging, IEEE Transactions on 33(1):135–146
[14] Haouchine N, Dequidt J, Peterlik I, Kerrien E, Berger MO, Cotin S (2013) Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. In: Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on, IEEE, pp 199–208
[15] Haouchine N, Cotin S, Peterlik I, Dequidt J, Lopez MS, Kerrien E, Berger MO (2015) Impact of soft tissue heterogeneity on augmented reality for liver surgery. Visualization and Computer Graphics, IEEE Transactions on 21(5):584–597
[16] Jiang J, Xing Y, Wang S, Liang K (2017) Evaluation of robotic surgery skills using dynamic time warping. Computer Methods and Programs in Biomedicine 152(Supplement C):71 – 83, DOI https://doi.org/10.1016/j.cmpb.2017.09.007, URL http://www.sciencedirect.com/science/article/pii/S0169260716308513
[17] Kazhdan M, Hoppe H (2013) Screened poisson surface reconstruction. ACM Trans Graph 32(3):29:1–29:13, DOI 10.1145/2487228.2487237, URL http://doi.acm.org/10.1145/2487228.2487237
[18] Kim JH, Bartoli A, Collins T, Hartley R (2012) Tracking by detection for interactive image augmentation in laparoscopy. Lecture Notes in Computer Science pp 246–255
[19] Klein G, Murray D (2007) Parallel tracking and mapping for small ar workspaces. In: Proc. 6th IEEE and ACM Int. Symp. Mixed and Augmented Reality, pp 225–234, DOI 10.1109/ISMAR.2007.4538852
[20] Knook M, Rosmalen A, Yoder B, Kleinrensink G, Snijders C, Looman C, Steensel C (2001) Optimal mesh size for endoscopic inguinal herniarepair. Surgical Endoscopy 15(12):1471–1477, DOI 10.1007/s00464-001-0048-9, URL https://doi.org/10.1007/s00464-001-0048-9
[21] Kratzer W, Fritz V, Mason RA, Haenle MM, Kaechele V, RSG (2003) Factors affecting liver size: a sonographic survey of 2080 subjects. J Ultrasound Med 22(11):1155–1161

[22] Kumar A, Wang YY, Wu CJ, Liu KC, Wu HS (2014) Stereoscopic visualization of laparoscope image using depth information from 3D model. Computer Methods and Programs in Biomedicine 113(3):862–868, DOI 10.1016/j.cmpb.2013.12.013, URL http://dx.doi.org/10.1016/j.cmpb.2013.12.013
[23] Levin D (2004) Mesh-independent surface interpolation. Mathematics and Visualization pp 37–49
[24] Lin B, Johnson A, Qian X, Sanchez J, Sun Y (2013) Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery. Lecture Notes in Computer Science pp 35–44
[25] Lin J, Clancy NT, Hu Y, Qi J, Tatla T, Stoyanov D, Maier-Hein L, Elson DS (2017) Endoscopic depth measurement and super-spectral-resolution imaging. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017
[26] London IC (2016) Hamlyn centre laparoscopic / endoscopic video datasets. URL http://hamlyn.doc.ic.ac.uk/vision/, [Accessed 6 Nov. 2016]
[27] Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2):91–110
[28] Mahmoud N, Cirauqui I, Hostettler A, Doignon C, Soler L, Marescaux J, Montiel J (2016) Orbslam-based endoscope tracking and 3d reconstruction. In: MICCAI 2016 CARE
[29] Michael K, Bolitho M, Hoppe H (2006) Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing, vol 7, p 2006
[30] Mistry M, Roach VA, Wilson TD (2013) Application of stereoscopic visualization on surgical skill acquisition in novices. Journal of Surgical Education 70(5):563 – 570, DOI http://dx.doi.org/10.1016/j.jsurg.2013.04.006, URL http://www.sciencedirect.com/science/article/pii/S1931720413001165
[31] Mountney P, Yang GZ (2008) Soft tissue tracking for minimally invasive surgery: learning local deformation online. Med Image Comput Comput Assist Interv 11(Pt 2):364–372
[32] Mountney P, Yang GZ (2009) Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society DOI 10.1109/iembs.2009.5333939, URL http://dx.doi.org/10.1109/IEMBS.2009.5333939
[33] Mountney P, Yang GZ (2010) Motion compensated slam for image guided surgery. Medical Image Computing and Computer-Assisted Intervention MICCAI 2010
[34] Mountney P, Stoyanov D, Yang GZ (2010) Three-dimensional tissue deformation recovery and tracking. IEEE Signal Processing Magazine 27(4):14–24, DOI 10.1109/msp.2010.936728, URL http://dx.doi.org/10.1109/MSP.2010.936728
[35] Mur-Artal R, Montiel JMM, Tardós JD (2015) Orb-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics 31(5):114–1163, DOI 10.1109/TRO.2015.2463671
[36] Murali A, Sen S, Kehoe B, Garg A, McFarland S, Patil S, Boyd WD, Lim S, Abbeel P, Goldberg K (2015) Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp 1202–1209, DOI 10.1109/ICRA.2015.7139344
[37] Nützi G, Weiss S, Scaramuzza D, Siegwart R (2011) Fusion of imu and vision for absolute scale estimation in monocular slam. Journal of intelligent & robotic systems 61(1):287–299
[38] Planteféve R, Peterlik I, Haouchine N, Cotin S (2016) Patient-specific biomechanical modeling for guidance during minimally-invasive hepatic surgery. Ann Biomed Eng 44(1):139–153, DOI 10.1007/s10439-015-1419-z, URL http://dx.doi.org/10.1007/s10439-015-1419-z
[39] Prados E, Faugeras O (????) Shape from shading: A well-posed problem? 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR?05) DOI 10.1109/cvpr.2005.319, URL http://dx.doi.org/10.1109/CVPR.2005.319
[40] Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: an efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, pp 2564–2571
[41] Saunders JA, Backus BT (2006) The accuracy and reliability of perceived depth from linear perspective as a function of image size. Journal of Vision 6(9):7–7
[42] Stoyanov D, Darzi A, Yang GZ (2004) Dense 3d depth recovery for soft tissue deformation during

robotically assisted laparoscopic surgery. Medical Image Computing and Computer-Assisted Intervention MICCAI 2004

[43] Stoyanov D, Darzi A, Yang GZ (2005) A practical approach towards accurate dense 3d depth recovery for robotic laparoscopic surgery. Comput Aided Surg 10(4):199–208, DOI 10.3109/10929080500230379, URL http://dx.doi.org/10.3109/10929080500230379

[44] Stoyanov D, Scarzanella MV, Pratt P, Yang GZ (2010) Real-time stereo reconstruction in robotically assisted minimally invasive surgery. Medical Image Computing and Computer-Assisted Intervention MICCAI 2010

[45] Su LM, Vagvolgyi BP, Agarwal R, Reiley CE, Taylor RH, Hager GD (2009) Augmented reality during robot-assisted laparoscopic partial nephrectomy: Toward real-time 3D-CT to stereoscopic video registration. Urology 73(4):896–900, DOI 10.1016/j.urology.2008.11.040

[46] Totz J, Mountney P, Stoyanov D, Yang GZ (2011) Dense surface reconstruction for enhanced navigation in mis. Med Image Comput Comput Assist Interv 14(Pt 1):89–96

[47] Velayutham V, Fuks D, Nomi T, Kawaguchi Y, Gayet B (2016) 3d visualization reduces operating time when compared to high-definition 2d in laparoscopic liver resection: a case-matched study. Surgical endoscopy 30(1):147–153

[48] Visentini-Scarzanella M, Stoyanov D, Yang GZ (2012) Metric depth recovery from monocular images using shape-from-shading and specularities. 2012 19th IEEE International Conference on Image Processing DOI 10.1109/icip.2012.6466786, URL http://dx.doi.org/10.1109/ICIP.2012.6466786

[49] Visentini-Scarzanella M, Sugiura T, Kaneko T, Koto S (2017) Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. International Journal of Computer Assisted Radiology and Surgery 12(7):1089, DOI 10.1007/s11548-017-1609-2, URL http://dx.doi.org/10.1007/s11548-017-1609-2

[50] Wagner OJ, Hagen M, Kurmann A, Horgan S, Candinas D, Vorburger SA (2012) Three-dimensional vision enhances task performance independently of the surgical method. Surg Endosc 26(10):2961–2968, DOI 10.1007/s00464-012-2295-3, URL http://dx.doi.org/10.1007/s00464-012-2295-3

[51] Wu CH, Sun YN, Chang CC (2007) Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. IEEE Transactions on Biomedical Engineering 54(7):1199–1211, DOI 10.1109/TBME.2006.889767

[52] Ye M, Giannarou S, Meining A, Yang GZ (2016) Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. Medical Image Analysis 30:14–157, DOI 10.1016/j.media.2015.10.003, URL http://dx.doi.org/10.1016/j.media.2015.10.003

[53] Zhang L, Ye M, Giannarou S, Pratt P, Yang GZ (2017) Motion-compensated autonomous scanning for tumour localisation using intraoperativeultrasound. Medical Image Computing and Computer-Assisted Intervention ? MICCAI 2017

[54] Zou Y, Liu PX (2017) A high-resolution model for soft tissue deformation based on point primitives. Computer Methods and Programs in Biomedicine 148(Supplement C):113 – 121, DOI https://doi.org/10.1016/j.cmpb.2017.06.013, URL http://www.sciencedirect.com/science/article/pii/S0169260716306071